Student Homework Sheet — Stage 05: Data Storage

Due: Next class session

Assignment

In the lecture, we learned **how to save/load CSV and Parquet with environment-driven paths, organize `data/raw/` vs `data/processed/`, and document storage choices**.
Now, you will adapt these patterns to implement a reproducible storage layer for one dataset.

Tasks

1. **Save in Two Formats**
    - Use a DataFrame (your own or provided sample).
    - Save to `data/raw/` as CSV and to `data/processed/` as Parquet.
    - Use `DATA_DIR_RAW` and `DATA_DIR_PROCESSED` from `.env`.
2. **Reload and Validate**
    - Reload both files.
    - Confirm shapes match and critical columns keep expected dtypes.
    - Add a small validation function and show results in the notebook.
3. **Refactor to Utilities**
    - Implement `write_df` and `read_df` that route by file suffix (csv/parquet).
    - Handle missing directories and missing Parquet engine with a clear message.
4. **Document**
    - Update `README.md` with a **Data Storage** section describing:
        - Folder structure (`data/raw/`, `data/processed/`)
        - Formats used and why
        - How your code reads/writes using env variables

Step-by-Step

- Start from `stage05_data-storage-preview_homework-starter.ipynb`.
- Fill TODOs: env paths, save/load, validation, utilities.
- Ensure `.env` contains:

```
DATA\_DIR\_RAW=data/raw
DATA\_DIR\_PROCESSED=data/processed
```

- Commit notebook and README changes; ensure folders exist.

Grading Rubric (100 pts)

- **Correct Save/Load (30)** — CSV + Parquet saved in right folders with env paths
- **Validation (20)** — Reload checks for shape/dtypes with clear output
- **Utilities (30)** — `write_df`/`read_df` robust to suffix/dirs/engine absence
- **Documentation (20)** — README storage section complete & clear

Example Deliverables

- `notebooks/` homework notebook (executed)
- `data/raw/sample_YYYYMMDD-HHMM.csv`
- `data/processed/sample_YYYYMMDD-HHMM.parquet`
- Updated `README.md`

> **Chain:** In the lecture, we learned **environment-driven IO, CSV vs Parquet, folder conventions, and validation**.
> **Now, you will adapt this** to **implement a reusable storage layer** with utilities and documentation.
